DOCUMENT RESUME

ED 341 734 TM 017 930

AUTHOR Bock, R. Darrell
TITLE The Graded Mark-Point Method of Scoring Performance Exercises and Open-Ended Items.
INSTITUTION Center for Research on Evaluation, Standards, and Student Testing, Los Angeles, CA.; Chicago Univ., Ill.; National Opinion Research Center, Chicago, Ill.
SPONS AGENCY Office of Educational Research and Improvement (ED), Washington, DC.
REPORT NO CSE-TR-323
PUB DATE Jun 91
CONTRACT OERI-G0086-003
NOTE 9p.
AVAILABLE FROM CSE Dissemination Office, UCLA Graduate School of Education, 405 Hilgard Avenue, Los Angeles, CA 90024-1521.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Educational Assessment; Evaluators; Grade 12; *Grading; High Schools; High School Seniors; *Item Response Theory; National Programs; Scaling; *Science Tests; *Scoring; Student Evaluation; Testing Programs
IDENTIFIERS *Graded Mark Point Scoring; Open Ended Questions; *Performance Based Evaluation

ABSTRACT

The scoring method that will be applied in the current 12th-grade science assessment project of the National Science Foundation and the Office of Educational Research and Assessment is described. The method, "graded mark-point" scoring, is modeled after procedures developed by P. Tamir for use in the performance exercises of the Israeli Matriculation Examination. The method is codified and made suitable for the item response theory scaling procedures that will be used in the analysis and reporting of assessment results. Development of the method has also been influenced by the scoring and scaling procedures of the California Direct-Writing Assessment, now used on a mass basis. Mark-point scoring uses rating forms that specify certain main points that the student should make in responding to the exercise or term. Points are identified by an expert who also provides one-sentence descriptions of the points as documentation on the rating form. Readers study the documentation and then, guided by the statements of each point, mark student papers for quality on each point. An example of the application of mark-point scoring to an open-ended item from the Earth Sciences section of the 12th-grade science assessment prototype is presented. Two figures illustrate the text. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED341734

C·R·E·S·S·T

Center for Research
on Evaluation, Standards,
and Student Testing

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

THE GRADED MARK-POINT METHOD OF SCORING PERFORMANCE EXERCISES AND OPEN-ENDED ITEMS

CSE Technical Report 323

R. Darrell Bock

The University of Chicago and NORC

UCLA Center for Research on Evaluation,
Standards, and Student Testing

TM017930

BEST COPY AVAILABLE

THE GRADED MARK-POINT METHOD OF SCORING PERFORMANCE EXERCISES AND OPEN-ENDED ITEMS

CSE Technical Report 323

R. Darrell Bock

The University of Chicago and NORC

UCLA Center for Research on Evaluation, Standards, and Student Testing

June 1991

The research reported herein was conducted with partial support from the U.S. Department of Education, Office of Educational Research and Improvement, pursuant to Grant No. G0086-003. However, the opinions expressed do not necessarily reflect the position or policy of this agency and no official endorsement by this agency should be inferred.

Please address inquiries to: CSE Dissemination Office, UCLA Graduate School of Education, 405 Hilgard Avenue, Los Angeles, California, 90024-1521

The increasing use of performance exercises and open-ended items in large-scale educational assessment has created a need for dependable and economical methods of scoring the student's responses. In assessment, the three main desiderata of any measurement procedure are (a) acceptable accuracy at the most detailed level of measurement, whether that of individual students, schools, programs, or larger entities, (b) consistency in stability over a period of years in the presence of changes in the exercises and items and of the reading teams who score their responses, and (c) sufficiently low unit costs to permit statewide or nationwide testing.

The present report describes the method of scoring that will be applied in the current NSF/OERI-supported twelfth-grade science assessment project (Bock & Doran, 1989). This method, which we refer to as "graded mark-point" scoring is modeled after procedures developed by Dr. Pinchas Tamir for use in the performance exercises of the Israeli Matriculation Examination. We have attempted to codify his method and make it suitable for the IRT scaling procedures that will be used in the analysis and reporting of assessment results. Development of the method has also been influenced by the scoring and scaling procedures of the California Direct-Writing Assessment, which are now employed on a mass basis. The following sections of this report describe the mark-point method, explain and justify the various steps in the procedure, and present an example of its application to an open-ended item from the Earth Sciences Section of the NSF/OERI twelfth-grade science assessment prototype.

The Scoring Method

Mark-point scoring makes use of rating forms, typically one 8 1/2 X 11 page, that specify certain main "points" that the student should make in responding to the exercise or item. The points should be identified by a person with expert knowledge of the topic in question (preferably the writer of the exercise); the expert must also supply documentation and commentary on the item for use in training the raters. The documentation should include one-sentence descriptions of the points for inclusion on the rating form. The readers who will rate the student responses must first study the documentation to familiarize themselves with the points the students are expected to make in each exercise; then, guided by the brief statements of each point on the corresponding rating form, they will look for each point in the student's paper and mark it for quality on the following six-point graded scale:

0 Point is not mentioned in paper.
1 Point is mentioned, but is incorrectly stated.
2 Point is mentioned, but is only partly correctly stated.
3 Point is mentioned, and is fully correctly stated.
4 Point is correctly stated and is partly elaborated.
5 Point is correctly stated and is fully elaborated.

The number of points to be rated depends on the type of exercise. For an elaborate laboratory exercise requiring perhaps 90 minutes of student time, the student's written record of hypotheses, observations, and conclusions might be rated on 16 to 20 distinct points. For a 10-minute response to an open-ended item of a paper-and-pencil test, rating of six to eight distinct points probably would be sufficient. A reader might be expected to spend 3 to 5 minutes in rating an open-ended item, but may require 10 to 15 minutes to rate the report of a 90-minute laboratory performance exercise.

Rationale of the Method

The mark-point method is designed to operate within the assumptions and limitations of large-scale assessment.

The performance exercises and open-ended items of the assessment are assumed to be problem-solving tasks. The main points to be looked for in the student's responses are their applications of scientific principles to the solution of the problem solved. They are not required, however, to use exact scientific terminology in their answers; it is assumed that knowledge of terminology is part of the multiple-choice section of the assessment instrument. The reader must attempt to infer the level of the student's understanding, in whatever language it is expressed.

This method of scoring is normative rather than descriptive. It does not attempt to classify the typical types of errors and misconceptions that students will inevitably make in responding to the novel situations presented in the exercises. Although such information may be of interest for some purposes, it is material for background research studies and not directly relevant to the evaluation goals of assessment.

The mark-point method is designed to make the rating procedure as objective as possible in order to achieve high levels of agreement between raters. It is the specific points that are to be rated and not an overall impression of the fluency or style of the paper. Some degree of subjectivity necessarily enters into the meaning of the graded categories of the rating scale, but it is assumed that by comparing their ratings of sample papers, the judges can attain reasonable levels of agreement. In an operational assessment that reports only at the school or higher levels, each student paper is read only once. Because papers from the same school are randomly assigned to reading team members, stability of the school mean score is attained by averaging the ratings of numerous readers. If the ratings are to be used in placement, advancement, or certification of individual students, however, more than one reading per paper would be desirable.

Because open-ended items require much more student time than multiple-choice items, it is important that a sufficient amount of information be extracted in the scoring process. This is the reason for using the graded scoring categories. When scaled by item response theoretic methods, a graded response typically has greater total information capacity than a multiple-choice item. Six to 8 such items can be equivalent to 15 or 20 multiple-choice items.

It is assumed that performance exercises and open-ended items evaluate different cognitive processes from multiple-choice items. For this reason it is important that they be scaled separately from multiple-choice items in the IRT analysis. If school-level scoring is assumed, separate scales for these items can be constructed even when each student responds to only one such item. For student-level scores, however, six to eight distinct items would be desirable. A relatively long testing time would be required for student-level measurement with open-ended items.

The weight that a particular point will receive in calculating a school-level or student-level scale score is determined by the IRT scaling procedure. It depends on the difficulty and discriminating power of the point and not on the arbitrary numbers zero through five that are used to label the rating categories.

Example of an Open-ended Item and Its Mark Points

An example of an open-ended item in Earth Science appears in Figure 1. The topic is lake effects on climate and weather. The mark points and possible elaborations for this item are as follows:

Principles

1. Winds in middle latitudes are prevailing westerly.
2. Large bodies of water warm and cool more slowly under radiant heating than land.

Seasonal Effects

1. City B will have cooler springs and warmer autumns than city A.
2. City B will have a longer frost-free period than city A.
3. City B will be subject to lake-effect snow in winter.
4. On clear, calm summer days following cloudless nights, both cities will experience land-to-lake breezes in early morning.
5. Under the same conditions, they will experience lake-to-land breezes in the afternoon.

Elaborations

1. Students may point out that the typical extremes of summer and winter temperatures in mid-continent, mid-latitude locations will be reduced by the presence of the lake, especially on its eastern shore.
2. Some students may know of the high specific heat of water, relative to earth, and of the effect of wind in mixing the waters of the lake in increasing its heat capacity.
3. Some may observe that we must assume the lake to be at a low enough altitude so as to not freeze over during the winter if lake-effect snows are to be expected on the eastern shore.
4. Observant students will also note that the lake is assumed to be deep enough, so that there is a sufficient volume of water to produce appreciable seasonal effects.

Mark-point Rating Form

The rating form for this item is shown in Figure 2. It is designed for reading by high-volume optical character recognition equipment.

Figure 1

Earth Science (Water)

The map below represents two cities on the shore of a large lake at 40° north latitude in the middle of a continent. How will the presence of the lake affect the climate of each of the cities at different times of the year and different times of the day? Give reasons for your answers.

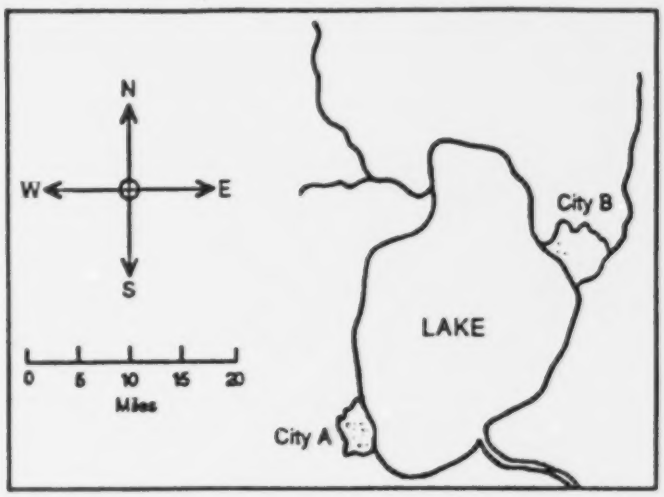

Figure 2

Open-ended item 22003002 Earth Science

School and Student ID No. ☐☐☐☐☐☐☐☐☐☐

Reader ID No. ☐☐☐☐☐☐☐☐☐☐

Date and Time ☐☐☐ ☐☐ : ☐☐

Topic: Lake Effect

Mark points

1. 0 1 2 3 4 5 ☐☐☐☐☐☐ "Winds prevailing westerly"
2. 0 1 2 3 4 5 ☐☐☐☐☐☐ "Large bodies of water warm and cool more slowly than land"
3. 0 1 2 3 4 5 ☐☐☐☐☐☐ "City B will have cooler springs and warmer autumns than City A"
4. 0 1 2 3 4 5 ☐☐☐☐☐☐ "City B will have a longer frost-free period than City A"
5. 0 1 2 3 4 5 ☐☐☐☐☐☐ "City B will be subject to lake-effect snow"
6. 0 1 2 3 4 5 ☐☐☐☐☐☐ "Both cities may experience land-to-lake breezes in mornings"
7. 0 1 2 3 4 5 ☐☐☐☐☐☐ "Both cities may experience lake-to-land breezes in afternoons"

Reader comments (optional):

Note: Use No. 2 pencil or erasable ball point pen.
Print number in block style: 0 1 2 3 4 5 6 7 8 9.
Mark rating boxes with an "X".
Erase all errors completely